

Combining Data-driven and Model-driven Methods for Robust Facial Landmark Detection

Hongwen Zhang, Qi Li, Zhenan Sun

Center for Research on Intelligent Perception and Computing

National Laboratory of Pattern Recognition

CAS Center for Excellence in Brain Science and Intelligence Technology

Institute of Automation, Chinese Academy of Sciences

hongwen.zhang@cripac.ia.ac.cn, {qli, znsun}@nlpr.ia.ac.cn

Abstract

Facial landmark detection is an important but challenging task for real-world computer vision applications. This paper proposes an accurate and robust approach for facial landmark detection by combining data-driven and model-driven methods. Firstly, a fully convolutional network (FCN) is trained to generate response maps of all facial landmark points. Such a data-driven method can make full use of holistic information in a facial image for global estimation of facial landmarks. Secondly, the maximum points in the response maps are fitted with a pre-trained point distribution model (PDM) to generate initial facial landmark shape. Such a model-driven method can correct the location errors of outliers by considering shape prior information. Thirdly, a weighted version of Regularized Landmark Mean-Shift (RLMS) is proposed to fine-tune facial landmark shapes iteratively. The weighting strategy is based on the confidence of convolutional response maps so that FCN is integrated into the framework of Constrained Local Model (CLM). Such an Estimation-Correction-Tuning process perfectly combines the global robustness advantage of data-driven method (FCN), outlier correction advantage of model-driven method (PDM) and non-parametric optimization advantage of RLMS. The experimental results demonstrate that the proposed approach outperforms state-of-the-art solutions on the 300-W dataset. Our approach is well-suited for face images with large poses, exaggerated expression, and occlusions.

1. Introduction

Facial landmark detection, which is also known as face alignment, aims to precisely and quickly localize a set of semantic points on a face image such as eye-corners, nose tip, and lips *etc.* It is a fundamental problem in computer vision

with wide applications in face recognition, facial expression analysis, human-computer interaction, video games *etc.* Great progress has been achieved in facial landmark detection and current methods can provide reliable results for near frontal face images [5, 23, 16, 21, 25]. But it is still a challenging problem for face images with partial occlusions or face images with appearance variations due to illumination, pose and expression changes.

Existing methods can be roughly divided into two categories according to their solution space, namely model-driven (parametric) methods and data-driven (non-parametric) methods.

Model-driven methods formulate the problem as a parametric model (e.g. Point Distribution Model [7] or PDM for short) and the objective is to search for the optimal parameters that best fit the evidence and prior. Active Appearance Model (AAM) [6], Active Shape Model (ASM) [8] and Constrained Local Model (CLM) [9] are some well known parametric model methods. AAM uses Principle Component Analysis (PCA) to generate a statistical model of shape and texture variations. Unlike AAM which uses a holistic appearance model, CLM models a set of local feature templates. Some patch experts are trained based on the local features. The landmark locations are obtained based on a joint optimization of local likelihood and global prior. CLM establishes a general framework of facial landmark detection and some improvements of patch experts and optimization method have been proposed in recent years [19, 2, 1].

In contrast, data-driven methods map the features extracted from face images to the facial landmark positions directly by learning a set of regressors on the training dataset. The landmark positions are typically updated iteratively using a coarse-to-fine strategy [23, 25]. And their regressors are typically builded in a cascaded manner [5, 16]. The shape constraint is naturally encoded into the cascaded regressors. The representative data-driven methods include

ESR [5], SDM [23] and deep learning based regression methods [21, 25]. ESR directly learns a regression function to infer the face shape from the training data. The errors between prediction and ground truth facial landmark positions are explicitly minimized in the learning procedure. SDM tries to learn a sequence of descend directions that minimize the facial landmark detection errors. Deep neural network as a typical data-driven pattern recognition method has been successfully applied to facial landmark detection in recent years. Deep Convolutional Network Cascade (DCNC) uses a three-level cascaded CNNs for facial landmark detection [21]. Coarse-to-fine Auto-Encoder networks (CFAN) uses cascaded Auto-Encoders with multi-scale inputs for accurate face alignment [25]. TCDCN uses only one CNN to achieve superior performance but it needs extra data with attribute label to train the tasks-constrained network [26, 27].

Both model-driven and data-driven methods are effective for locating facial landmarks on the near frontal face images. However neither method is perfect to handle challenging facial samples and each method has its specific advantages and disadvantages. For example, model-driven methods typically constrain the facial shape or appearance in a predefined model. However facial images in the wild usually have large shape or appearance variations which do not follow the models. Therefore model-driven methods may crash in handling these samples because of their representation limitation in capturing complex and subtle face variations. Data-driven methods (*e.g.*, SDM, LBF) proposed in recent years are more flexible to alleviate the problems mentioned above. However, the data-driven methods also suffer from local minimum when starting from a less accurate initial shape. So both data-driven and model-driven methods are complementary in facial landmark detection and it is better to make full use of their advantages and avoid their disadvantages.

This paper proposes a novel framework namely ECT (Estimation-Correction-Tuning) for facial landmark detection by combining data-driven and model-driven methods. The novelty and contributions of this paper are summarized as follows.

(a) Data-driven estimation of initial landmarks: A fully convolutional network (FCN) is proposed to learn a desirable response map for each landmark. The ideal response map is defined as a 2D Gaussian and its center is the facial landmark. Such a data-driven method can make full use of holistic information to avoid local minimum traps.

(b) Model-driven correction of outliers: The initial landmarks extracted from the response maps are fitted into a pre-trained Point Distribution Model (PDM), so that the outlier landmarks in data-driven method can be corrected using prior information of facial shape model. Both the expression power of data-driven method (deep CNN) and the

reasonable shape constraints (PDM) are unified into a general framework for robust facial landmark detection.

(c) Non-parametric fine-tuning of facial shape: The confidence of the response map is integrated into a weighted version of Regularized Landmark Mean-Shift (RLMS) [19] framework. The combination of the evidence from the response maps and parametric shape prior improves facial landmark detection further. Utilizing the shape prior is beneficial to infer the positions of the invisible landmarks in occlusion cases.

(d) The proposed ECT method achieves state-of-the-art performance on the 300-W dataset. The success of the Estimation-Correction-Tuning strategy and the idea of combining data-driven and model-driven methods finds a new way to develop more advanced solutions for robust facial landmark detection. Moreover, our method is also highly potential for other CVPR problems such as human pose estimation, image segmentation, *etc.*

The reminder of this paper is organized as follows. Section 2 introduce technical details of the proposed approach. Experimental results are reported in Section 3. Finally, Section 4 concludes this paper.

2. Approach

The flowchart of the proposed ECT (Estimation-Correction-Tuning) approach is shown in Fig. 1. There are mainly three steps namely Estimation Step, Correction Step and Tuning Step to achieve a robust facial landmark detection result given an input face image. The Estimation Step aims to obtain a global estimation of the initial landmarks based on the maximum response points on the response maps, which are learned from a fully convolutional network (FCN). The Correction Step aims to achieve a more reasonable initial shape by correcting the outlier landmarks using a pre-trained point distribution model (PDM). The Tuning Step aims to fine-tune the landmark shape based on a weighted version of Regularized Landmark Mean-Shift (RLMS).

This Section introduces problem formulation firstly for understanding the framework of our approach. And then the significant parts of the ECT approach namely convolutional response map and weighted regularized mean shift are presented respectively.

2.1. Problem formulation

Point distribution model (PDM) [7] is widely used in classic parametric methods. It models the shape with global rigid transformation (scaling, in-plane rotation and translation) and non-rigid shape variations (head poses and expressions). For a 2D shape $\mathbf{x} = [\mathbf{x}_1; \dots; \mathbf{x}_n] \in \mathbb{R}^{2n \times 1}$, where $\mathbf{x}_i = [x_i; y_i]$ denotes the 2D coordinate of the i^{th} landmark,

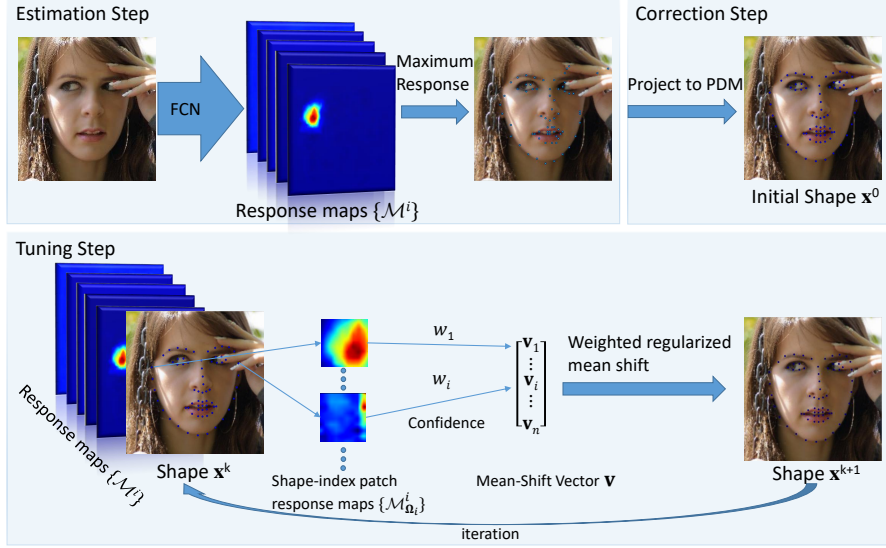


Figure 1: Flowchart of the proposed ECT (Estimation-Correction-Tuning) approach for facial landmark detection.

it can be expressed in the following equation:

$$\mathbf{x}_i = s\mathbf{R}(\bar{\mathbf{x}}_i + \Phi_i \mathbf{q}) + \mathbf{t} \quad (1)$$

where $\mathbf{p} = \{s, \mathbf{R}, \mathbf{t}, \mathbf{q}\}$ denotes the PDM parameters, which consist of a set of global rigid transform parameters (global scaling s , rotation \mathbf{R} , translation \mathbf{t}) and a set of non-rigid parameters \mathbf{q} . $\bar{\mathbf{x}}_i$ and Φ_i denote the pertaining sub-matrix in the mean shape $\bar{\mathbf{x}}$ and the shape eigenvectors Φ respectively. Φ is the eigenvectors corresponding to the m largest eigenvalues by applying PCA to a set of training shapes. Given enough training samples, Φ can code the rich expression adequately. Assuming the rigid transformation has a non-informative prior and the nonrigid shape parameter \mathbf{q} exhibits the Gaussian distribution, the PDM parameter \mathbf{p} has the following prior:

$$p(\mathbf{p}) \propto \mathcal{N}(\mathbf{q}; \mathbf{0}, \mathbf{\Lambda}); \quad \mathbf{\Lambda} = \text{diag} \{[\lambda_1; \dots; \lambda_m]\} \quad (2)$$

where λ_i denotes the eigenvalue of the i^{th} eigenvector in Φ . The parameter \mathbf{p} can be inferred in a Bayesian manner. Assuming the detections are conditionally independent for each landmark, the posterior distribution of \mathbf{p} can be written as:

$$p(\mathbf{p} | \{l_i = 1\}_{i=1}^n, \mathcal{I}) \propto p(\mathbf{p}) \prod_{i=1}^n p(l_i = 1 | \mathbf{x}_i, \mathcal{I}) \quad (3)$$

where $l_i \in \{1, -1\}$ denotes whether the i^{th} landmark is aligned or misaligned on coordinate \mathbf{x}_i for image \mathcal{I} .

Given the detection expert \mathcal{D}_i for the i^{th} landmark, the likelihood $p(l_i = 1 | \mathbf{x}_i, \mathcal{I})$ takes the form:

$$p(l_i = 1 | \mathbf{x}_i, \mathcal{I}) \propto p(\mathcal{D}_i | \mathcal{I}) p(l_i = 1 | \mathbf{x}_i, \mathcal{D}_i, \mathcal{I}) \quad (4)$$

where $p(\mathcal{D}_i | \mathcal{I})$ can be regarded as the confidence of the detection expert \mathcal{D}_i for image \mathcal{I} , and $p(l_i = 1 | \mathbf{x}_i, \mathcal{D}_i, \mathcal{I})$ is the conditional likelihood corresponding to the align probability of the i^{th} landmark under the context of using detect expert \mathcal{D}_i . Assuming that there are a set of candidate coordinates Ψ_i for the i^{th} landmark, $p(l_i = 1 | \mathbf{x}_i, \mathcal{D}_i, \mathcal{I})$ can be approximated with a nonparametric representation [19] of Kernel Density Estimator (KDE). So the conditional likelihood has the following form:

$$p(l_i = 1 | \mathbf{x}_i, \mathcal{D}_i, \mathcal{I}) = \sum_{\mathbf{y}_i \in \Psi_i} \pi_{\mathbf{y}_i} \mathcal{N}(\mathbf{x}_i; \mathbf{y}_i, \rho_i \mathbf{I}) \quad (5)$$

where $\pi_{\mathbf{y}_i} = p(l_i = 1 | \mathbf{y}_i, \mathcal{D}_i, \mathcal{I})$ is corresponding to the response map (normalized) introduced in the next subsection. $\rho_i = \frac{\rho^2}{w_i}$ is used to smooth the response map, where ρ is a free parameter and $w_i = p(\mathcal{D}_i | \mathcal{I})$ adjusts the smoothness.

Combining Eqs. (3)(4) and (5), we can get the following form:

$$\begin{aligned} p(\mathbf{p} | \{l_i = 1\}_{i=1}^n, \mathcal{I}) \\ \propto p(\mathbf{p}) \prod_{i=1}^n p(\mathcal{D}_i | \mathcal{I}) \sum_{\mathbf{y}_i \in \Psi_i} \pi_{\mathbf{y}_i} \mathcal{N}(\mathbf{y}_i; \mathbf{x}_i, \rho_i \mathbf{I}) \end{aligned} \quad (6)$$

Eq. (6) can be solved iteratively using the EM algorithm and the mean-shift algorithm [19]. In the E-step, the posterior over \mathbf{y}_i can be evaluated as follows when the candidates $\{\mathbf{y}_i\}_{i=1}^n$ are regarded as hidden variables:

$$w_{\mathbf{y}_i} = p(\mathbf{y}_i | l_i = 1, \mathbf{x}_i, \mathcal{D}_i, \mathcal{I}) = \frac{\pi_{\mathbf{y}_i} \mathcal{N}(\mathbf{x}_i; \mathbf{y}_i, \rho_i \mathbf{I})}{\sum_{\mathbf{z}_i \in \Psi_i} \pi_{\mathbf{z}_i} \mathcal{N}(\mathbf{x}_i; \mathbf{z}_i, \rho_i \mathbf{I})} \quad (7)$$

Then, the M-step involves minimizing the Q function:

$$\begin{aligned} Q(\mathbf{p}) &= E_{q(\mathbf{y})}[-\ln\{p(\mathbf{p}) \prod_{i=1}^n p(l_i = 1, \mathbf{y}_i | \mathbf{x}_i, \mathcal{I})\}] \\ &\propto \|\mathbf{q}\|_{\Lambda^{-1}}^2 + \sum_{i=1}^n w_i \sum_{\mathbf{y}_i \in \Psi_i} \frac{w_{\mathbf{y}_i}}{\rho^2} \|\mathbf{x}_i - \mathbf{y}_i\|^2 \end{aligned} \quad (8)$$

where $q(\mathbf{y}) = \prod_{i=1}^n p(\mathbf{y}_i | l_i = 1, \mathbf{x}_i, \mathcal{D}_i, \mathcal{I})$. The iterative solution $\Delta \mathbf{p}$ for each update can be written as:

$$\Delta \mathbf{p} = -(\rho^2 \tilde{\Lambda}^{-1} + \mathbf{J}^T \mathbf{W} \mathbf{J})^{-1} (\rho^2 \tilde{\Lambda}^{-1} \mathbf{p} - \mathbf{J}^T \mathbf{W} \mathbf{v}) \quad (9)$$

where $\tilde{\Lambda} = \text{diag}\{[0; \lambda_1; \dots; \lambda_m]\}$, and $\mathbf{J} = [\mathbf{J}_1; \dots; \mathbf{J}_n]$ where \mathbf{J}_i is the Jacobian of PDM in Eq. (1), $\mathbf{W} = \text{diag}\{[w_{x_1}; w_{y_1}; \dots; w_{x_n}; w_{y_n}]\}$ with $w_{x_i} = w_{y_i} = w_i$, $\mathbf{v} = [\mathbf{v}_1; \dots; \mathbf{v}_n]$ is the mean shift vectors of all landmarks:

$$\mathbf{v}_i = \left(\sum_{\mathbf{y}_i \in \Psi_i} \frac{\pi_{\mathbf{y}_i} \mathcal{N}(\mathbf{x}_{\mathbf{y}_i}^c; \mathbf{y}_i, \rho_i \mathbf{I})}{\sum_{\mathbf{z}_i \in \Psi_i} \pi_{\mathbf{z}_i} \mathcal{N}(\mathbf{x}_{\mathbf{z}_i}^c; \mathbf{z}_i, \rho_i \mathbf{I})} \mathbf{y}_i \right) - \mathbf{x}_i^c \quad (10)$$

where \mathbf{x}_i^c is the currently estimated position of the i^{th} landmark.

The main difference between our formulation and RLMS [19] is that a weight is assigned to each landmark mean-shift vector and it is projected onto the subspace spanned by the PDM's Jacobian, which contributes a key factor to the success in robust facial landmark detection. Eqs. (9) and (10) indicate that our algorithm alternates between computing the move step from response maps and regularizing it with the shape model's constraint.

2.2. Convolutional response map

Data-driven methods in literature [5, 21, 23] learn a regression of landmark location $[x_i; y_i]$ directly. The novelty of our solution is to regress an ideal response map for each landmark using FCN. The ideal response map of the i^{th} landmark for image \mathcal{I} is a gray image \mathcal{M}^i with the same size of \mathcal{I} , and its pixel value on position \mathbf{z} can be defined as $\mathcal{M}_{\mathbf{z}}^i = \mathcal{N}(\mathbf{z}; \mathbf{x}_i^*, \sigma^2 \mathbf{I})$, where \mathbf{x}_i^* is the ground truth location of the i^{th} landmark, and σ is used to control the scope of the response.

Fig. 2 shows an overview of the proposed FCN architecture. Our network consists of three connected subnetworks, namely PrimaryNet, FusionNet, and UpsampleNet. Given an input image \mathcal{I} with the size of 256×256 , the first two subnetworks regress the smaller response maps with the size of 64×64 . The last subnetwork is a deconvolution network which stacks the smaller response maps together and upsamples them to the size of \mathcal{I} . The details of filter parameters can be found in the supplementary materials.

Given the training dataset $N = \{(\mathcal{I}, \mathbf{x}^*)\}$, where \mathbf{x}^* is the ground truth shape of image \mathcal{I} , the training objective

of the regressor becomes the task of estimating the network weights λ that minimize the following L2 loss function:

$$l(\lambda) = \sum_{(\mathcal{I}, \mathbf{x}^*) \in N} \sum_i \|\mathcal{M}^i - \phi^i(\mathcal{I}, \lambda)\|^2 \quad (11)$$

where $\phi^i(\mathcal{I}, \lambda)$ is the i^{th} channel output of the regression network fed with image \mathcal{I} . The loss functions for PrimaryNet and FusionNet are just the smaller scale version of (11).

The origin of PrimaryNet and FusionNet is from the pose estimation networks [15]. And it is adapted to the case of facial landmark detection. As mentioned in [15], PrimaryNet can not learn the spatial dependencies of landmarks very well. So conv3 and conv7 are concatenated firstly to address this problem as suggested in [15]. And then they are fed to FusionNet. Such a strategy generates satisfactory response maps from FusionNet. And the output of PrimaryNet is only slightly inferior to FusionNet since the loss of FusionNet is also backpropagated through PrimaryNet. The main architecture difference between our subnetworks and the pose estimation networks [15] is that the layers of conv4, conv5, conv1_fusion, conv2_fusion and conv3_fusion are divided into several sub-layers which use dilated convolution kernel [24] instead of dense convolution kernel. All these original layers are replaced by the new version of the concatenated sub-layers as shown in Fig. 2. Such an improvement can achieve a comparable regression result with 20 megabytes of model size reduced, which accounts for more than 40 percentage of the whole model size in our case. The benefit of dilated convolution is significant, as it supports exponential expansion of the receptive field with the number of parameters growing linearly. These dilated convolution kernels are well-suited to landmark detection task which requires the pixel level texture information in multiple scales.

2.3. Weighted regularized mean shift

During testing, the image \mathcal{I} is firstly fed into a pre-trained FCN to obtain response maps $\mathcal{M} = \{\mathcal{M}^i\}$. In the initialization period, the facial landmarks are first located with the maximum response positions in response maps. Then the coarse result is projected onto the PDM space to obtain a reasonable initial landmark shape \mathbf{x}^0 . And then the estimated shape is fine-tuned iteratively using weighted regularized mean shift.

In the k^{th} stage, the currently estimated shape is \mathbf{x}^k and the shape-index patch response maps are extracted from \mathcal{M} for each landmark. The shape-index patch response map $\{\pi_{\mathbf{y}_i}\}$ for the i^{th} landmark is a $r \times r$ square subarea centered at \mathbf{x}_i^k in the response map \mathcal{M}^i . For those areas out of the range of \mathcal{M}^i (e.g. \mathbf{x}_i^k is close to the boundary of \mathcal{M}^i when r is large enough), they are padded with zeros. All coordinates in the square subarea for the i^{th} landmark are

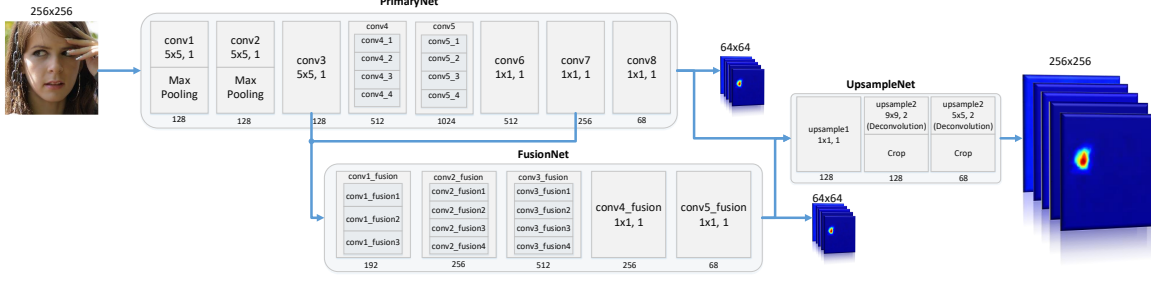


Figure 2: Architecture of the proposed fully convolutional network (FCN).

denoted as Ω_i . Inspired by the work in [16], the patch size r is shrunk from early stage to later stage.

For each shape-index patch response map, its confidence w_i is estimated empirically as follows:

$$w_i = \text{sigmoid}(a \frac{\sum_{\mathbf{y}_i \in \Omega_i} \pi_{\mathbf{y}_i}}{\text{var}_{\mathbf{y}_i \in \Omega_i}(\pi_{\mathbf{y}_i})} + b) \quad (12)$$

where $\text{var}_{\mathbf{y}_i \in \Omega_i}(\pi_{\mathbf{y}_i})$ denotes the variance of the patch response map, a and b are two empirical parameters which can be optimized via cross-validation. Eq. (12) suggests that the confidence of a local expert is proportional to its response value and inversely proportional to the dispersed degree of its spatial distribution. Sigmoid function is used to constrain the confidence value at the range of $(0, 1)$.

Then the mean shift vectors \mathbf{v}_i are computed for all landmarks using Eq. (10), where the candidate coordinates Ψ_i are set with Ω_i , and the response map $\{\pi_{\mathbf{y}_i}\}$ is normalized so that $\sum_{\mathbf{y}_i \in \Psi_i} \pi_{\mathbf{y}_i} = 1$. The confidence w_i is assigned to each mean shift vector \mathbf{v}_i , and the update $\Delta \mathbf{p}$ for PDM parameters can be computed with the following projection in Eq. (9). Finally the latest estimated shape \mathbf{x}^{k+1} is obtained for the next iteration by applying the incremental version of Eq. (1).

The fine-tuning process can be converged to stable results after a number of iterations. A desirable result can be achieved after 10 iterations in the experiments. The complete process of the algorithm is summarized in Algorithm 1.

3. Experiment

Both training and testing experiments of the proposed method are conducted on the 300-W dataset [18], which is the most widely used dataset for robustness evaluation of facial landmark detection algorithms. The 300-W dataset contains face images re-annotated from the existing datasets, including LFPW [3], AFW [29], HELEN [13], XM2VTS [14], and 135 images in challenging poses and expressions (called iBUG dataset). FCN and PDM are

Algorithm 1 ECT (Estimation-Correction-Tuning) for facial landmark detection.

Require:

The pre-trained PDM and FCN, the test image \mathcal{I}

Output:

The final shape \mathbf{x}

▷ **Estimation step**

- 1: Feed \mathcal{I} in FCN to obtain the response maps $\{\mathcal{M}^i\}$
- 2: Obtain the coarse shape by locating the landmarks with the maximum response positions in $\{\mathcal{M}^i\}$

▷ **Correction step**

- 3: Project the coarse shape onto the PDM and we get the initial shape \mathbf{x}^0

▷ **Tuning step**

- 4: **for** $k = 0$ to K **do**
- 5: **for** the i^{th} landmark **do**
- 6: Calculate the shape-index patch coordinates Ω_i
- 7: Extract the patch response map $\{\pi_{\mathbf{y}_i} = \mathcal{M}_{\mathbf{y}_i}^i\}$ where $\mathbf{y}_i \in \Omega_i$
- 8: Calculate \mathbf{v}_i and w_i from $\{\pi_{\mathbf{y}_i}\}$, using Eqs. (10) and (12) respectively
- 9: **end for**
- 10: Calculate $\Delta \mathbf{p}$ using Eq. (9)
- 11: Update the PDM parameters $\mathbf{p}^{k+1} = \mathbf{p}^k + \Delta \mathbf{p}$
- 12: Update the shape \mathbf{x}^{k+1} using Eq. (1)
- 13: **end for**

trained on 3,148 images from the whole AFW, the training set of LFPW, and the training set of HELEN. The testing dataset includes 689 images from the test set of LFPW, the test set of Helen, and the whole iBUG. Viola-Jones face detector is used to get the initial face region for each training and testing image. The size of face region is doubled before rescaling it to 256×256 .

3.1. Implementation details

We implement FCN using the Caffe framework [11]. The FCN takes the input of a 256×256 face image, and outputs a set of response maps with the same resolution of 256×256 . To avoid overfitting, we randomly flip the input image horizontally and randomly crop a 248×248 sub-image from it. Then, we rotate it with a random angle from -30° to 30° before rescaling it back to 256×256 . The variance σ of the 2D Gaussians for the final response maps is set to 6. For the smaller response maps generated from PrimaryNet and FusionNet, we set σ to 1.5. The network training is divided into two stages. We first train PrimaryNet and FusionNet, and then use the trained model to fine-tune the whole FCN (the layer learning rates of the first two subnetworks are set to 0). The learning rate at the first stage is set to 10^{-6} , and it is decreased to 10^{-7} at 20K iterations. The learning rate at the second stage is fixed to 10^{-8} . The momentums in both stages are set to 0.95.

3.2. Evaluation on the 300-W dataset

Experimental results are reported on the 300-W dataset. The evaluation on the 300-W consists of three parts, namely the common subset, the challenging subset and the full set. The common subset includes 554 images from the test set of LFPW and HELEN, and the challenging subset contains 135 images from iBUG. The full set uses all of them with 689 images in total. For fair comparison, the localization error is normalized by the inter-pupil distance in consistency with [16, 28]. The evaluation metric is the mean error, which can be calculated as $\frac{1}{n} \sum_{i=1}^n \frac{\|\mathbf{x}_i - \mathbf{x}_i^*\|^2}{d}$, where d denotes the inter-pupil distance, and n is the number of facial landmarks. Since the pupil landmark position are not available in the 300-W dataset, it is estimated by averaging the coordinates of the landmarks around the eye [16, 28]. The cumulative error distribution (CED) curve is also used to evaluate the localization performance.

The methods used for comparison include Zhu *et al.* [29], Smith *et al.* [20], DRMF [2], GN-DPM [22], RCPR [4], CFAN [25], ESR [5], SDM [23], ERT [12], LBF [16], CFSS [28] and TCDCN [26]. The mean errors of different methods are reported in Table 1. Performance degrade is observed for all methods on the challenging subset iBUG in comparison with the results on the common subset. And the comparison shows that the proposed ECT method outperforms all other state-of-the-art methods on the fullset. Our method is slightly inferior to TCDCN on the challenging subset (mean error 8.69 vs. 8.60). It should be noted that TCDCN uses additional training data labeled with facial attributes. We plot the cumulative error distribution (CED) curves of different methods in Figure 3. As shown in Figure 3, the performance of ECT is superior to other methods particularly on the challenging subset, which means ECT is

robust to various challenging conditions (*e.g.*, exaggerated expression, occlusion) as shown in Figure 4 and Figure 5.

Method	Common Subset	Challenging Subset	Fullset
Zhu <i>et al.</i> [29]	8.22	18.33	10.20
Smith <i>et al.</i> [20]	-	13.30	-
DRMF [2]	6.65	19.79	9.22
GN-DPM [22]	5.78	-	-
RCPR [4]	6.18	17.26	8.35
CFAN [25]	5.5	16.78	7.69
ESR [5]	5.28	17.00	7.58
SDM [23]	5.57	15.4	7.50
ERT [12]	-	-	6.40
LBF [16]	4.95	11.98	6.32
CFSS [28]	4.73	9.98	5.76
TCDCN [26]	4.80	8.60	5.54
ECT	4.68	8.69	5.47

Table 1: Comparison of mean error with state-of-the-art methods on 300-W.

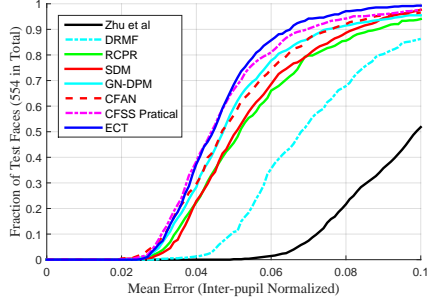
Experiment	Common Subset	Challenging Subset	Fullset
CRM(Baseline)	5.66	10.85	6.68
CRM-PDM	5.31	10.27	6.28
CRM-RLMS	4.84	9.64	5.78
ECT	4.68	8.69	5.47

Table 2: Success factor analysis of key components in our method.

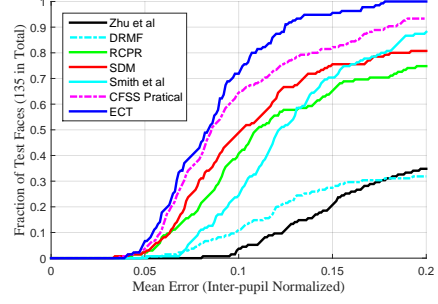
3.3. Success factor analysis

It is interesting to investigate the contribution of each module of the proposed ECT method. In this subsection, we analyse the components of ECT on the 300-W dataset. The results are shown in Table 2. CRM denotes the baseline method which simply locates the landmarks with the maximum response positions in the response maps. The subsequent methods denoted in Table 2 are the variant methods based on CRM, which will be introduced shortly.

Benefit from PDM. CRM-PDM is a simple combination of CRM and PDM. The improvement is significant on both the common subset and the challenging subset in the comparison between CRM and CRM-PDM. The results demonstrate the successful strategy of combing data-driven (CRM) and model-driven (PDM) methods.



(a) CED for 68-pts common subset of 300-W



(b) CED for 68-pts challenging subset of 300-W

Figure 3: Comparisons of cumulative errors distribution (CED) curves on 300-W.

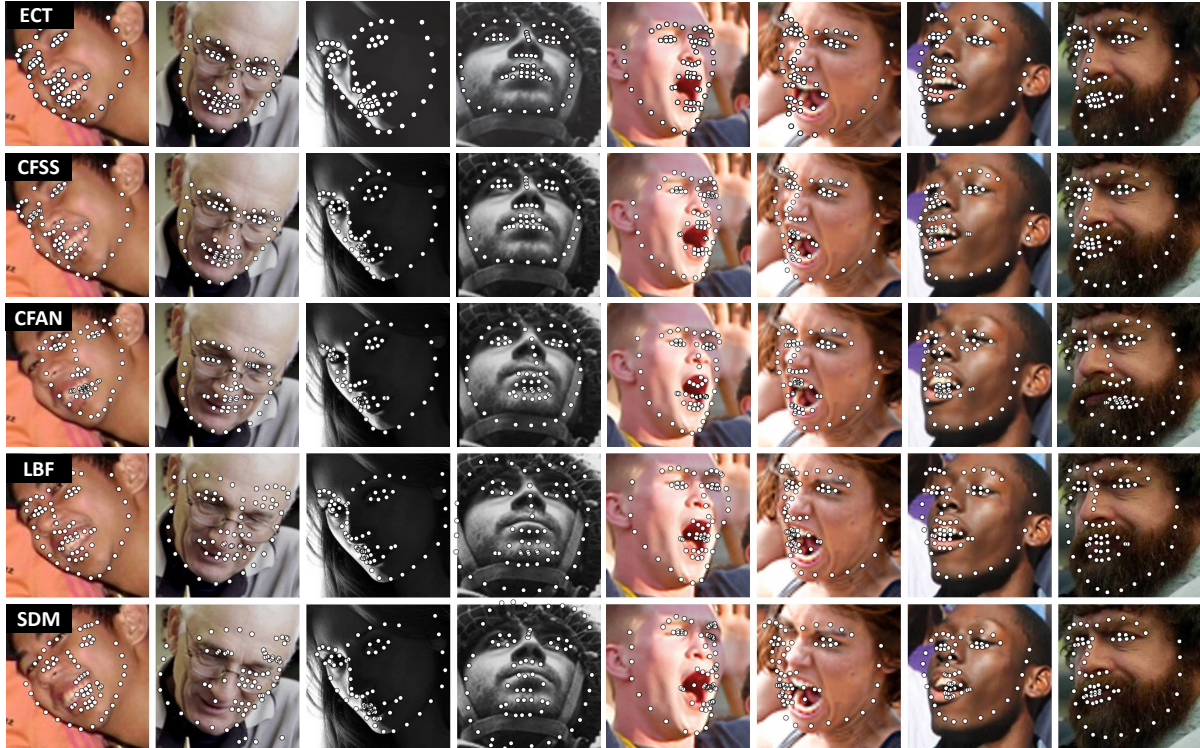


Figure 4: Example results for comparison of state-of-the-art methods. The results of other methods are from [28].

Benefit from weighted regularized mean shift. A comparison between ECT and CRM-PDM shows that the tuning step based on weighted regularized mean shift is necessary to achieve a better performance. The mean errors can be decreased by 11.9% and 15.4% on the common subset and the challenging subset, respectively. Joint optimization of response maps and PDM in weighted regularized mean shift framework leads to significant improvement in the challenging subset. The improvement comes from not only the prior information for inferring the invisible landmarks, but also the weighted mean shift vectors. To verify the conclusion,

the weights in Eq. (10) are removed, which is equivalent to setting all weights as ones. This approach can be regarded as the original Regularized Landmark Mean-Shift (RLMS) with its local experts extracted from our response maps. Such a method is denoted as CRM-RLMS and its results are reported on the third row in Table 2. The results show that CRM-RLMS is better than CRM-PDM but worse than the proposed ECT method. So the proposed weighting strategy in regularized mean shift is successful because the confidence of the response maps provides a more reasonable balance between the prior and the evidence.

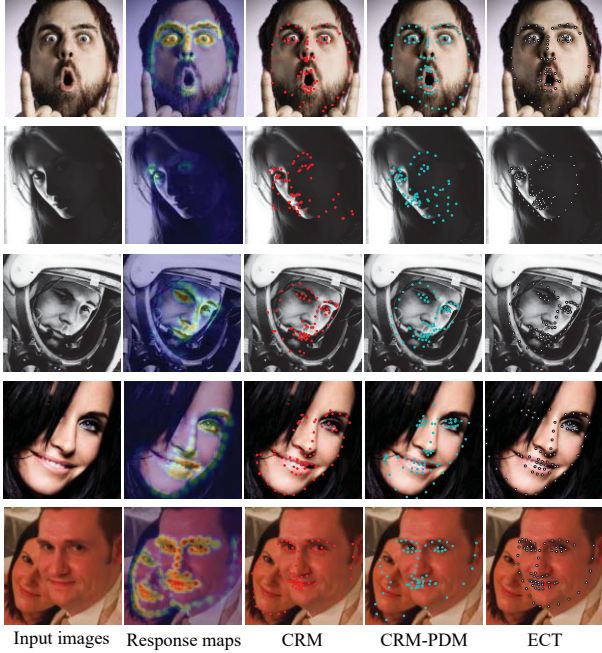


Figure 5: Results of the components of our ECT method. The images shown in column 1 are the original images. The response maps stacked with the original image together are shown in column 2. The results of CRM are shown in column 3. The outputs of CRM-PDM are shown in column 4. The outputs of the proposed ECT method are shown in the last column.

Furthermore, Figure 5 shows the facial landmark detection results of the three significant components of our ECT on several challenging images of 300-W. The results illustrate that the proposed ECT can infer the location of invisible landmarks better with the guidance of PDM prior and response maps.

3.4. Further investigation on the response maps

The response maps produced by FCN are invariant to translation, scale and rotation considerably. The response maps of several face images are shown in Figure 5. It is observed that the response maps are robust against large head pose, expression, and scale variations in face images. The invariance of the response maps makes great contributions to the robustness of our method to large pose and exaggerated expression. Shape initialization from such desirable response maps can make full use of the holistic information so that our method is not prone to local minimum. To verify the conclusion, an artificial disturbance is added to the bounding boxes provided by the face detector. Specifically, for both x and y directions, we jitter the center of bounding box randomly within 10% of the box width, and enlarge or

Experiment	Common Subset	Challenging Subset	Fullset
ECT	4.68	8.69	5.47
ECT+disturb	4.74	8.84	5.54

Table 3: Robustness test against face detection variations.

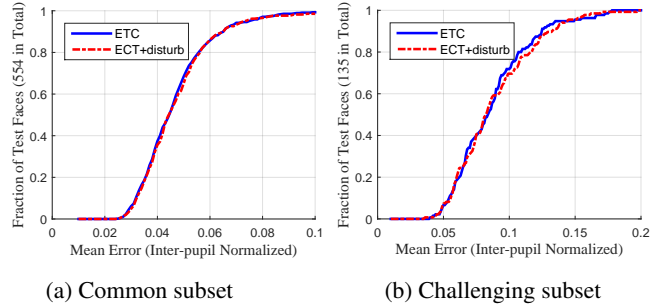


Figure 6: CED curve of our method on the 300-W dataset before and after disturbing the face detector.

shrink the bounding box randomly within 20% of the box width. Table 3 and Figure 6 show the results of our method before and after disturbing the face detector. It can be observed that the performance of our method degrades slightly but it still achieves state-of-the-art performance. Hence, our method is friendly to different face detectors [17] and is well suited to practical applications.

4. Conclusion

Both model-driven and data-driven methods have been proposed for facial landmark detection in the literature. They both have some specific limitations. This paper proposes a three-step (Estimation-Correction-Tuning) framework to combine model-driven and data-driven methods for robust facial landmark detection. The proposed ECT method achieves the outstanding results in comparison of state-of-the-art methods on the 300-W dataset. The success of the method comes from a good design of convolutional response maps (CRM), a good fusion of fully convolutional network (FCN), Point Distribution Model (PDM) and Regularized Landmark Mean-Shift (RLMS) and a good flowchart of Estimation-Correction-Tuning (ECT). And the novel ideas proposed in this paper such as ECT, CRM, CRM-PDM, CRM-RLMS, weighted RLMS, *etc.* are applicable to many similar problems in computer vision and pattern recognition. In the future, we will investigate ECT further in the context of general object alignment, human pose estimation, and image segmentation, *etc.*

References

- [1] J. Alabort-i Medina and S. Zafeiriou. Unifying holistic and parts-based deformable model fitting. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3679–3688, 2015. **1**
- [2] A. Athana, S. Zafeiriou, S. Cheng, and M. Pantic. Robust discriminative response map fitting with constrained local models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3444–3451, 2013. **1, 6**
- [3] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2930–2940, 2013. **5**
- [4] X. P. Burgos-Artizzu, P. Perona, and P. Dollár. Robust face landmark estimation under occlusion. In *IEEE International Conference on Computer Vision*, pages 1513–1520, 2013. **6**
- [5] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. *International Journal of Computer Vision*, 107(2):177–190, 2014. **1, 2, 4, 6**
- [6] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):681–685, 2001. **1**
- [7] T. F. Cootes and C. J. Taylor. Active shape models smart snakes. In *British Machine Vision Conference*, pages 266–275. 1992. **1, 2**
- [8] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59, 1995. **1**
- [9] D. Cristinacce and T. F. Cootes. Feature detection and tracking with constrained local models. In *British Machine Vision Conference*, volume 1, page 3, 2006. **1**
- [10] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst, 2007. **10**
- [11] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM international conference on Multimedia*, pages 675–678, 2014. **6**
- [12] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874, 2014. **6**
- [13] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In *European Conference on Computer Vision*, pages 679–692, 2012. **5**
- [14] K. Messer, J. Matas, J. Kittler, J. Luetttin, and G. Maitre. Xm2vtsdb: The extended m2vts database. In *Second international conference on audio and video-based biometric person authentication*, volume 964, pages 965–966, 1999. **5**
- [15] T. Pfister, J. Charles, and A. Zisserman. Flowing convnets for human pose estimation in videos. In *IEEE International Conference on Computer Vision*, pages 1913–1921, 2015. **4**
- [16] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 fps via regressing local binary features. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1685–1692, 2014. **1, 5, 6**
- [17] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment via regressing local binary features. *IEEE Transactions on Image Processing*, 25(3):1233–1245, 2016. **8**
- [18] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *IEEE International Conference on Computer Vision Workshops*, pages 397–403, 2013. **5**
- [19] J. M. Saragih, S. Lucey, and J. F. Cohn. Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision*, 91(2):200–215, 2011. **1, 2, 3, 4**
- [20] B. M. Smith, J. Brandt, Z. Lin, and L. Zhang. Nonparametric context modeling of local appearance for pose-and expression-robust facial landmark localization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1741–1748, 2014. **6**
- [21] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3476–3483, 2013. **1, 2, 4**
- [22] G. Tzimiropoulos and M. Pantic. Gauss-newton deformable part models for face alignment in-the-wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1851–1858, 2014. **6**
- [23] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 532–539, 2013. **1, 2, 4, 6**
- [24] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. **4**
- [25] J. Zhang, S. Shan, M. Kan, and X. Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *European Conference on Computer Vision*, pages 1–16. 2014. **1, 2, 6**
- [26] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *European Conference on Computer Vision*, pages 94–108. 2014. **2, 6**
- [27] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Learning deep representation for face alignment with auxiliary attributes. *IEEE transactions on pattern analysis and machine intelligence*, 38(5):918–930, 2016. **2**
- [28] S. Zhu, C. Li, C. Change Loy, and X. Tang. Face alignment by coarse-to-fine shape searching. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4998–5006, 2015. **6, 7**
- [29] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2879–2886, 2012. **5, 6**

Appendix

A. Supplementary results

Figure 7 shows more results to supplement Figure 5. Figure 8 shows selected results for partial face images cropped from the LFW dataset [10]. Note that we do not re-train our FCN on the LFW dataset, which indicates that the proposed method has a good generalization ability and can even deal with the case of partial faces in the wild.

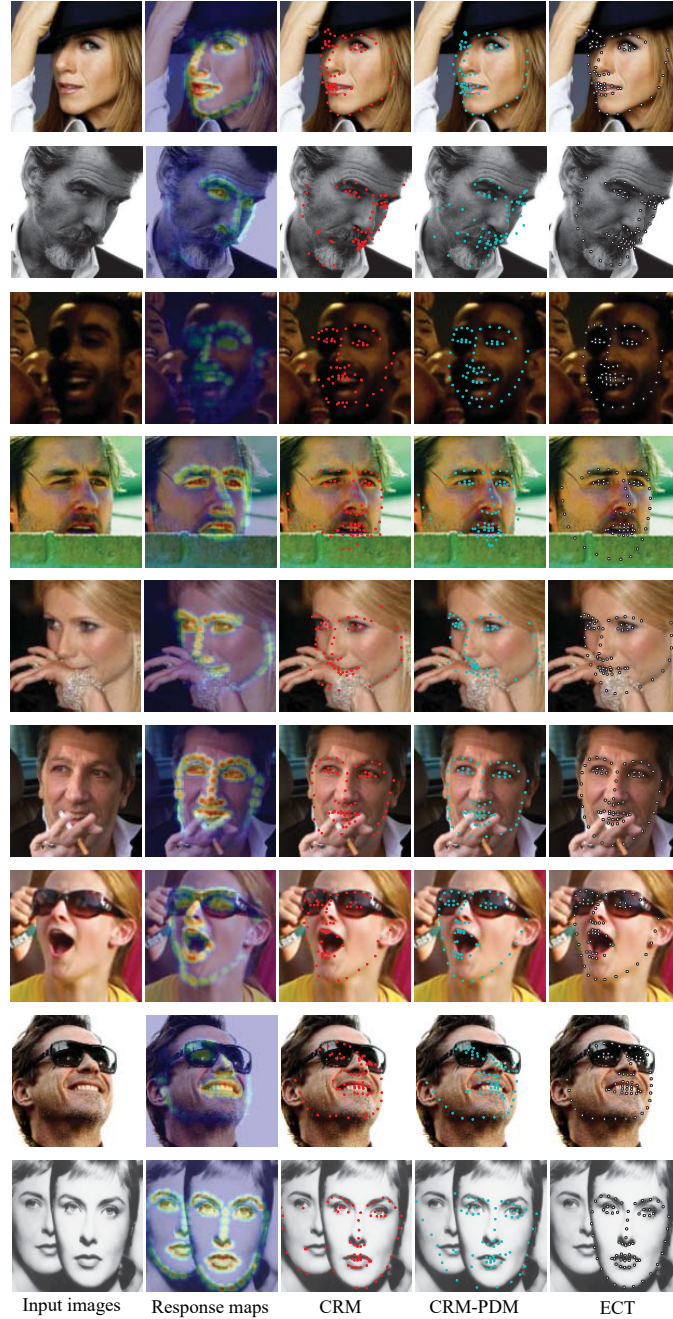


Figure 7: Supplementary results of the components in our ECT method.

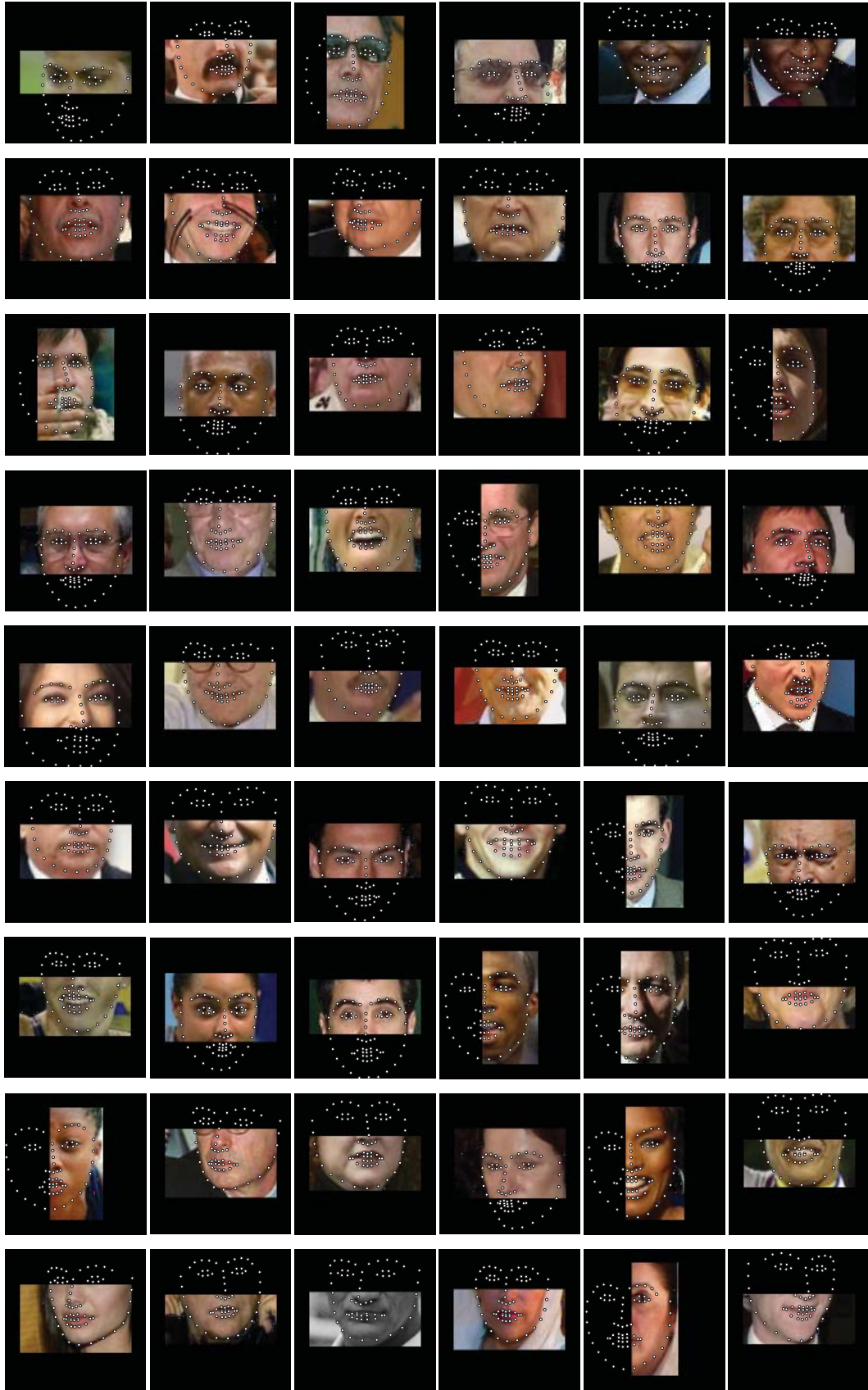


Figure 8: Selected results for partial face images cropped from the LFW dataset.